13.12.2012

Implementacja własnego algorytmu scoringu w Lucene

Dominika Puzio

Przypadek testowy

Problem:

Wyszukiwarka obrazków:

- system pozwala użytkownikom na wyszukiwanie obrazków po ich metadanych:
  - tekstowych: nazwa, opis, tagi
  - liczbowych: proporcja (szerokość/wysokość)
- użytkownik podaje szukane słowo oraz zakres proporcji, jaki go interesuje (np. od 1,5 do 2)
- system ma zwrócić obrazki ze znalezionym słowem w którymś z pól tekstowych (uwzględniając standardową trafność) i dodatkowo ocenić jakość dopasowania proporcji obrazka do zadanego przedziału

Przykład:

Obrazki w systemie:

nazwa: *lucene*
opis: *logo Lucene*
proporcja: *6,5*

nazwa: *luke*
opis: *logo Luke'a – czytnika indeksów Lucene*
proporcja: *1,4*

nazwa: *solr*
opis: *logo Solra – serwera wyszukiwawczego opartego na Lucene*
proporcja: *1,8*

Zapytanie:

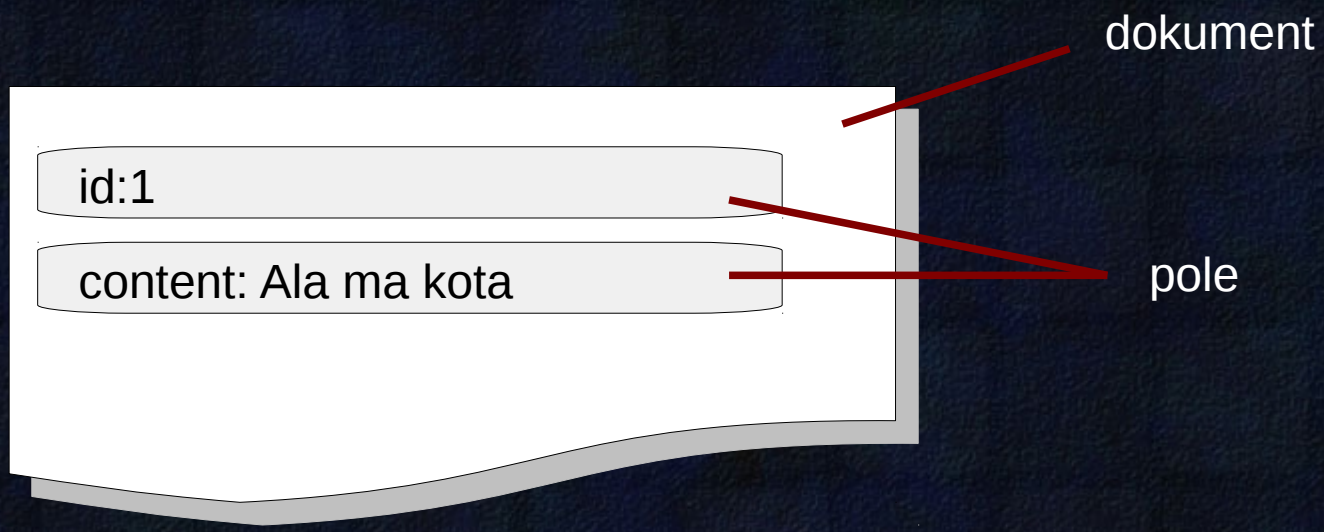
słowo: *Lucene* AND zakres proporcji: *[1,5 – 2,0]*

czyli szukamy obrazków o proporcjach od [ ] *do* [ ]

Podstawy teoretyczne: dokument

Dokument:

jednostka danych, pojedynczy element na liście wyników wyszukiwania, to co chcemy wyszukiwać (strona www, artykuł, post, dobro konsumpcyjne)



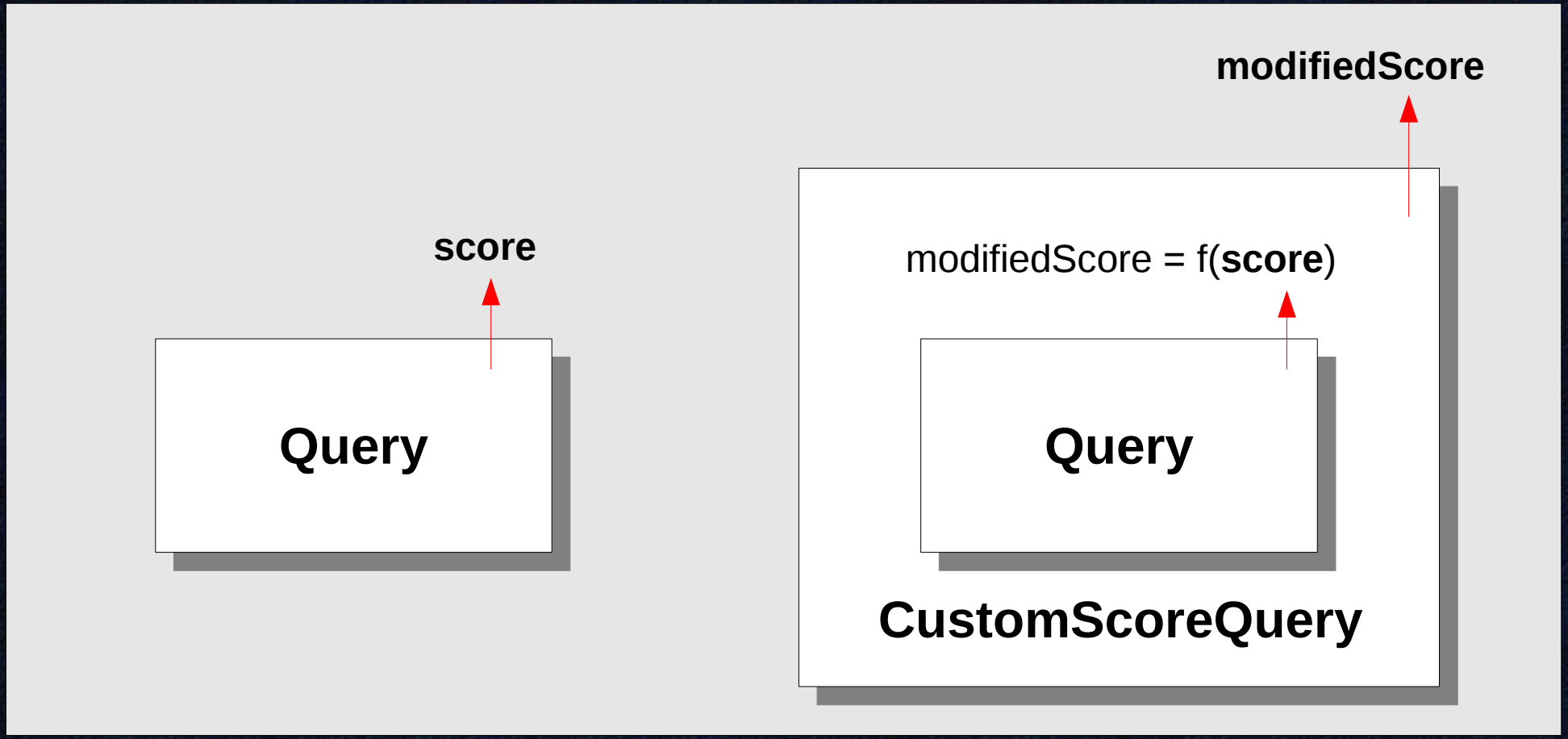
Problem:

Rozwiązanie:

Modyfikacja algorytmu scoringu Lucene, tak, aby oprócz standardowych kryteriów brał pod uwagę również odległość proporcji obrazka od żądanego przedziału.

CustomScoreQuery

- **CustomScoreQuery** opakowuje inne zapytanie, dając dostęp do jego *score'a* i pozwalając go zmienić



Indeks odwrócony

doc1

nazwa: lucene
opis: logo Lucene
proporcja: 6.5

doc2

nazwa: luke
opis: logo Luke'a
proporcja: 1.4

doc3

nazwa: solr
opis: logo Solra
proporcja: 1.8

nazwa:lucene	1
nazwa:luke	2
nazwa:solr	3
opis:logo	1 2 3
opis:lucene	1
opis:luke'a	2
opis:solra	3
proporcja:1.4	2
proporcja:1.8	3
proporcja:6.5	1

FieldCache

Słownik w indeksie odwróconym

Nie nadaje się do wyciągania informacji o wartości konkretnego pola w konkretnym dokumencie.

FieldCache

Tablica z wartościami pola dla wszystkich dokumentów.

- jedna wartość dla każdego dokumentu
- osobny FieldCache dla każdego pola

FieldCache dla pola *nazwa:*

lucene	luke	solr

FieldCache dla pola *proporcja:*

6.5	1.4	1.8